UNCLASSIFIED

AD. 297 279

Reproduced
by the

ARMED SERVICES TECHNICAL INFORMATION AGENCY
ARLINGTON HALL STATION
ARLINGTON 12, VIRGINIA

63-2-5.

AN APPLIED RESEARCH PROGRAM OF INTELLIGENCE DATA HANDLING

QUARTERLY REPORT NO. 1

May 1, 1961 - July 31, 1961

Contract AF 19 (626)-10

Submitted

to

by

INTERNATIONAL BUSINESS MACHINES CORPORATION
Thomas J. Watson Research Center
P. O. Box 218
Yorktown Heights, New York

AN APPLIED RESEARCH PROGRAM OF
INTELLIGENCE DATA HANDLING

QUARTERLY REPORT NO. 1

Task 1 --- Automatic Print Reading

A continuing monitoring effort is being maintained with respect to any automatic print reading devices or methods which may prove to be worthy of incorporation into a data handling system. No funding support has been requested from the Air Force for this monitoring effort. Personnel concerned with this effort are as follows:

M. Clayton Andrews
L. P. Horwitz
G. W. King
Glenn Shelton, Jr.

Task 2 --- Storage and Addressing

The method of integration of the 7090 and the AN/GSQ-16 Photostore should be both economical and practical. From the design point of view, the most economical and practical way of achieving this integration would be through the use of the standard interface channel. A standard interface would be put into the photostore. The photostore could then be connected to any 7090 that had a standard interface channel.

Any method of integrating these two units must be examined not only from an engineering point of view, but from an operational point of view as well. Some of the hardware and how it is connected depends upon how the photostore is to be operated. For example; All of the search logic of the present photostore would not be retained. The amount of search logic retained, if any, would depend upon how the photostore was to be operated.

The operation of the photostore with various levels of search logic is under investigation. As more is learned about the type of data to be stored on the disc and how this data is to be used, then the exact manner of operation can be determined.

Personnel presently performing full-time on this task are:

F. E. Franklin

Other personnel being assigned to this task are:

D. Caplan	1/4 time
E. P. Clarke	1/4 time
W. B. Strohm	1/4 time
R. D. Villani	1/4 time
Miss Janet Andress	1/4 time

Task 3 --- Multiple Interrogation

A brief discussion will ensue indicating the general concepts of the supervisory and control program. In particular the activity flow and the prime executive section of this program will be described. There exist many critical areas in which information is quite scanty, hence a brief telephone conversation with Lt. E. Morenoff had transpired Wednesday, August 16, 1961. As a result of this conversation, a definite meeting was established for Friday, September 1, 1961, to convene at IBM Research. Hopefully, many of the problem gaps may be clarified.

I would like to state what we mean by multiprocessing and multiprogramming within the frame-work of computer jargon:

Multiprocessing is the simultaneous execution of two or more usually independent and predetermined sequence of tasks, performed within a well defined organizational environment.

Multiprogramming can be, ideally considered as the concurrent activity of two or more sequences of instructions by and within computer environment, in which the respective flow of activities do not jeopardize any other within this environment.

The supervisory and control program will be a collection of subprograms, most of which will reside in cores, and will be the heart of the proposed IBM 7090 multiprocessing system. This program will perform a great many tasks and the following functions necessary to achieve a satisfactory solution as indicated and desired by the terms of the contract. These functions consist of:

1. Memory allocation schemes (implied relocation and memory protection).

2. The formation and resolution of priority work queues.

3. Investigations, initiation and control of all I-0 request and interrupts

4. Procedural time accounting of respective tasks

5. Inter-computer communication and linkages

6. Ascertain degree of diagnostic evaluation

7. Development of desired statistical information.

To perform most of the functions stated above, it is apparent that the fundamental task of the supervisor is to direct and maintain an orderly process, by which various tasks may be performed by operational programs and computer components (especially input-output). This process can be easily administered by means of several chained lists. We will define a list as a collection of items which possess a defined successor or predecessor or both relationship. Each list will have a name. An item is defined as a sequential block of core storage or elements. The elements are either data or program directives. The optimum size of an item will be determined and fixed when a more comprehensive notion is attained concerning the overall system parameters and requirements. Some items will be utilized to retain I-0 data and/or operational programs. Other items will be employed to serve as inter-computer communication center linkages.

It is evident that additions to any of the list are initiated and imposed by input-output demands as dictated by operational and control programs. Conversely, items are deleted from lists upon completion and satisfaction of an operational program or the termination of any output component, activity. Hence, the control or supervisor program must have the ability of directing activities according to some pre-determine hierarchy of priorities and to effectively create the proper implementation, upon an interpretive scheme, additions and deletions to the various lists. Briefly this implies the development of set of macros, which will at compile time describe a set of parameters indicating some intelligible action to be rendered. An example may be , take $I_1 \ T_1 \ K$ which may be interpreted as an assignment of N-item blocks.

Since it is assumed, that the collection of operational programs will far exceed the capacity of the 7090 system, means will be provided to have them placed on a high capacity, quick access file (Photo-disc store or 1301 file). Consequently, a programming philosophy must be revolved, so that programmers utilizing the system may be guided and adhere to restrictions and requirements so specified. Programs are on an external storage device that can be easily and quickly called upon request. One has several means by which to read them into high speed memory. Programs can be compiled in either absolute or relocatable form. Since program areas may be assigned anywhere in available high speed memory it is assumed that all programs will be compiled in relocatable form. Also, to avoid an enormous time delay penalty in program relocation, there must be a suitable hardware relocation scheme devised to undertake this task automatically.

A memory protection scheme should be considered. IBM product line has such a scheme, for various block-item size, which will be investigated.

A time accounting procedure is to be formalized as to the method of time sharing costs, while operating in the multiprocessing state. There are several viewpoints concerning this concept which will be investigated.

Notes, diagrams and flowcharts concerning the general planning of the IBM 7090 multiprocessing system for this contract are being maintained and updated. These will be illustrated and aired at the September 1, 1961 meeting with Lt. E. Morenoff. Hopefully, with the acquisition of additional information and distillation of primary goals, a more concise and direct presentation of the system will evolve.

Personnel assigned to this task are as follows:

R. D. Villani	3/4 time
Miss Janet Andress	3/4 time

Task 4 --- Correlation Techniques

1. Classification and Distribution of Data.

Suppose that it is required to assign an object, characterized by the values of p measurements, to one of m classes. For the case where the p measurements are continuous random variables, and the m p-dimensional distributions corresponding to these classes are known, the Neyman-Pearson lemma applies, and rules for assigning an object to a class have been explicitly stated and programmed by P. D. Welch. (1960). If the m distributions are each assumed to be multivariate normal, both with and without equal covariance matrices, procedures for estimating the population means and using these for classifying the object have been obtained.

Similar results for the case where some of this p measurements in the m populations are discrete random variables have been found in the classical statistical literature and investigated for application to the classification problem. For example, an optimum rule for qualitative classification was found, the effect of initial sample size and of p on the misclassification probability was computed. All of these results are for the special case of $m = 2$.

IBM 7090 test programs which take as input repeated samples six measurements on spoken vowels, have in storage some statistical data about the various distributions, and as output, signal the classification of a sound. These programs can handle up to 50 variables for any object but an unlimited number of observations. In 8 minutes of computer running time, 16 out of 20 objects were correctly classified. This program has been generalized and is complete now.

These results and the underlying techniques are presented in detail in Appendix C of "Classification Program For Aerospace Intelligence Data System", 31 March 1961, IBM FSD report, by C. Abraham.

Since March 31, 1961 the following research has been accomplished.

1. Reduction of variables by step-wise regression and we can prove that this procedure gives an optimum rule for reduction of variables in classification .

2. If in the mixed variable situation, the categorical variables are dicotomons in nature and their effect on the continuous variables is simply a shift in the mean value, then the point-biserial correlations between the two types of variables can be defined and optimum classification procedures obtained.

3. In the multivariate case, the effect of correlation on the "distance function" has been investigated and some general principles for inclusion or exclusion of in- dividual variables have been obtained.

4. A classification procedure without a strictly defined hypothesis about population parameters is under investigation.

These results are being written up as a technical report.

2. Lattice Theory Studies (An Experimental System for the Exchange of Scientific Information: DICO)

Since the last report, the following progress connected with DICO can be reported. A paper describing the entire system, in- corporating its design, the theoretical considerations underlying it and some preliminary experimental results was written and is being submitted for publication (Kochen and Wong). A copy of this paper

is enclosed as an appendix. Considerable thought has been given to an improved definition, procedure for testing and use of a person's professional interests as the fundamental variable in this system. Recommendations concerning new sampling methods as well as some deeper pertinent sociological considerations are forthcoming shortly. Consultation by R. Walker, a professional sociologist). Some novel ideas for modifying and extending work on DICO were proposed and analyzed in a preliminary way by M. Flood (Consultant).

3. Searching Models

In computer applications involving the storage of a large volume of information, frequent use is made of storage units characterized by one or both of the following restrictions:

1. Immediate access in the strict sense is not possible.

2. The precise address of an arbitrary item in storage is not specified a priori.

Fundamental to the operations of such memories are good strategies of search. As an initial step in the study of optimum search strategies, an one-dimensional model has been constructed and the search problem formulated. While the similarity between this theoretical model and the magnetic tape operations suggests itself immediately, the results apply as well to one of the dimensions of multidimensional storage devices, e.g., track search for magnetic and photo disks.

Assuming that the basic operations available are continuous search and skip, a class of search strategies have been obtained. The average time of search for these strategies is always less than that of continuous search (sequential reading) of the storage unit from beginning to end.

Recent developments include a re-formulation of the problem using dynamic programming techniques. One major result of this reformulation has been the proof that the class of strategies mentioned in the last paragraph are not optimal in the true sense, in that they only yield a lower bound of the mean search time and not the minimum.

Starting with the equations for optimization in the dynamic programming formulation present plans call for the following studies in the immediate future:

1. to obtain proof for some general theorems on the nature of optimum solutions, e.g., conditions for existence and uniqueness, etc.

2. to obtain the optimum solutions in some specific cases, e.g., uniform distribution.

3. failing in the analytical approach, to solve the optimization problem recursively, for which the dynamic programming approach is ideally suited. Such recursive operations are naturally adapted for use on the read-only memories, of which the photo-store is the major example.

The following personnel have been assigned to various projects under this Task on a part-time basis:

C. T. Abraham
M. Kochen
Miss P. Reisner
E. Wong

The following personnel will also contribute to some of these projects but will not be charged to the contract:

Miss P. B. Baxendale
B. Dunham
G. W. King
R. W. Moss
A. L. Samuel